CAPSTONE PROJECT

EDA ON HOTEL BOOKING ANALYSIS

BY
CHETAN BHANGARE
NIKHIL BHANGARE

PROBLEM STATEMENT

1. For this project we will be analyzing Hotel Booking data. This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces.
2. Hotel industry is a very volatile industry and the bookings depends on above factors and many more.
3. The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management ,which can perform various campaigns to boost the business and performance.

- We will divide our work flow into three steps

Data Collection and Understanding	Data Cleaning and Manipulation	Exploratory Data Analysias(EDA)

- EDA will be divided into following 3 UBM rule analysis.

- 1) Univariate analysis: Univariate analysis is the simplest of the three analyses where the data
- you are analyzing is only one variable.
- 2) Bivariate analysis: Bivariate analysis is where you are comparing two variables to study their
- relationships.
- 3) Multivariate anlysis: Multivariate analysis is similar to Bivariate analysis but you are
- comparing more than two variables

Data Collection and Understanding

- After collecting data it’s very important to understand your data. So we had hotel Booking analysis data.
- Which had 119390 rows and 32 columns. So let’s understand this 32 columns.
- Data Description:
- hotel :Resort Hotel or City Hotel
- is_canceled : Value indicating if the booking was canceled (1) or not (0)
- lead_time : Number of days that elapsed between the entering date of the booking and the arrival date

Data Collection and Understanding

- arrival_date_year : Year of arrival date
- arrival_date_month : Month of arrival date
- arrival_date_week_number : Week number of year for arrival date
- arrival_date_day_of_month : Day of arrival date
- stays_in_weekend_nights : Number of weekend nights
- stays_in_week_nights : Number of week nights.
- adults : Number of adults
- children : Number of children
- babies : Number of babies
- meal : Type of meal booked.
- country : Country of origin.

Data Collection and Understanding

- market_segment : Market segment designation. (TA/TO)
- distribution_channel : Booking distribution channel.(T/A/TO)
- is_repeated_guest : is a repeated guest (1) or not (0)
- previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking
- previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking
- reserved_room_type : Code of room type reserved.
- assigned_room_type : Code for the type of room assigned to the booking.
- booking_changes : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation

Data Collection and Understanding

- deposit_type : No Deposit, Non Refund , Refundable.
- agent : ID of the travel agency that made the booking
- company : ID of the company/entity that made the booking .
- days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer
- customer_type : type of customer. Contract,Group,transient,Transient party.
- adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- required_car_parking_spaces : Number of car parking spaces required by the customer
- total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)
- reservation_status : Reservation last status

Data Cleaning

```
duplicate_rows_df = df[df.duplicated()].shape

print(f"the no. of duplicate rows :" , duplicate_rows_df)
```

```
the no. of duplicate rows : (31994, 32)
```

Lets drop the duplicate values

```
df=df.drop_duplicates()
df.shape
```

```
(87396, 32)
```

Data Cleaning

```
df.isnull().sum()
```

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           452
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           12193
company                         82137
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

Since the column named **Company and Agents** have lots of null values , we will drop these columns

Data Cleaning

```
df = df.drop(columns=['company','agent'])
```

```
df.isnull().sum()
```

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           452
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

Since we only have children and country data with null values , so drop the unavailable data

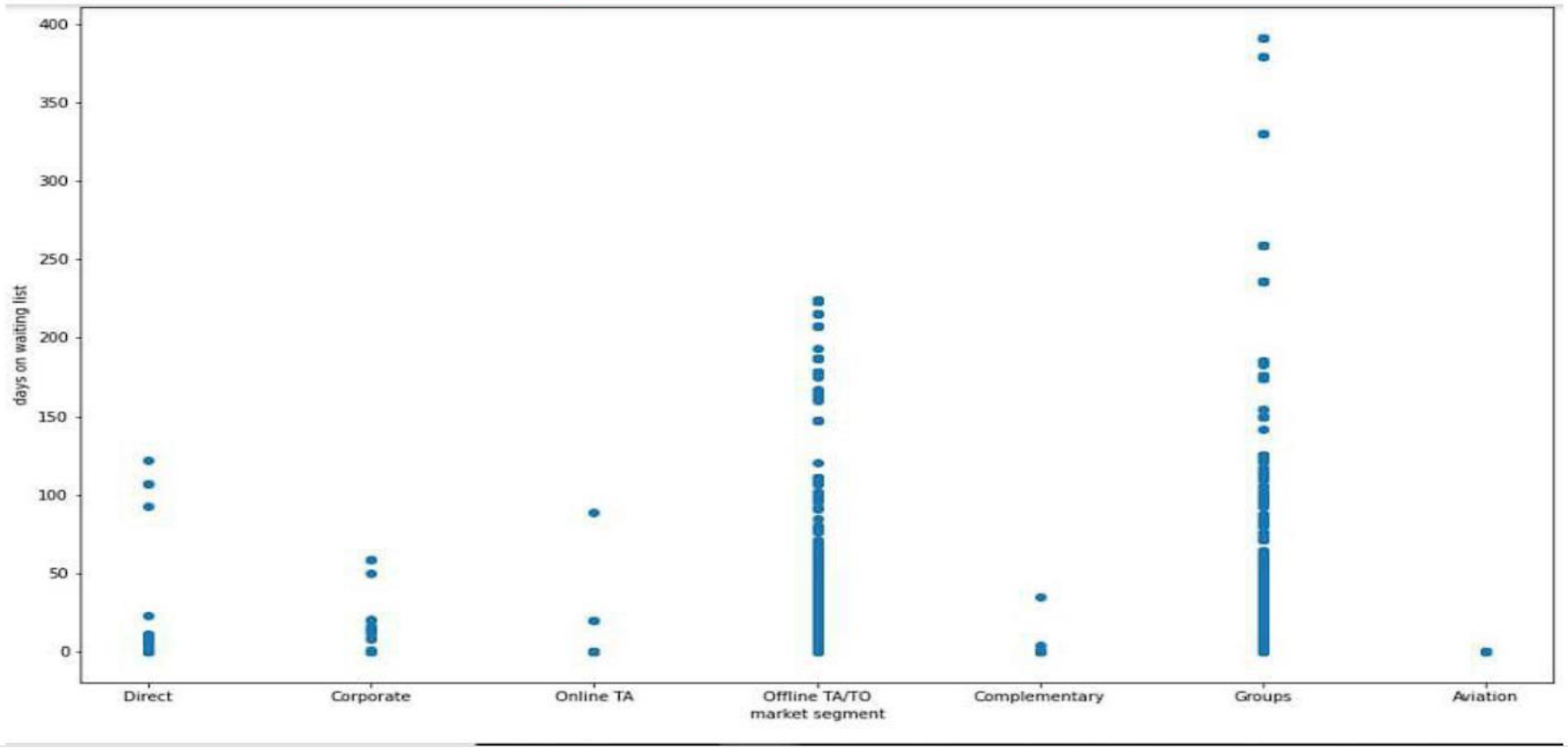
```
df=df.dropna()

#To insure we don't have any null values
df.isnull().sum()
```

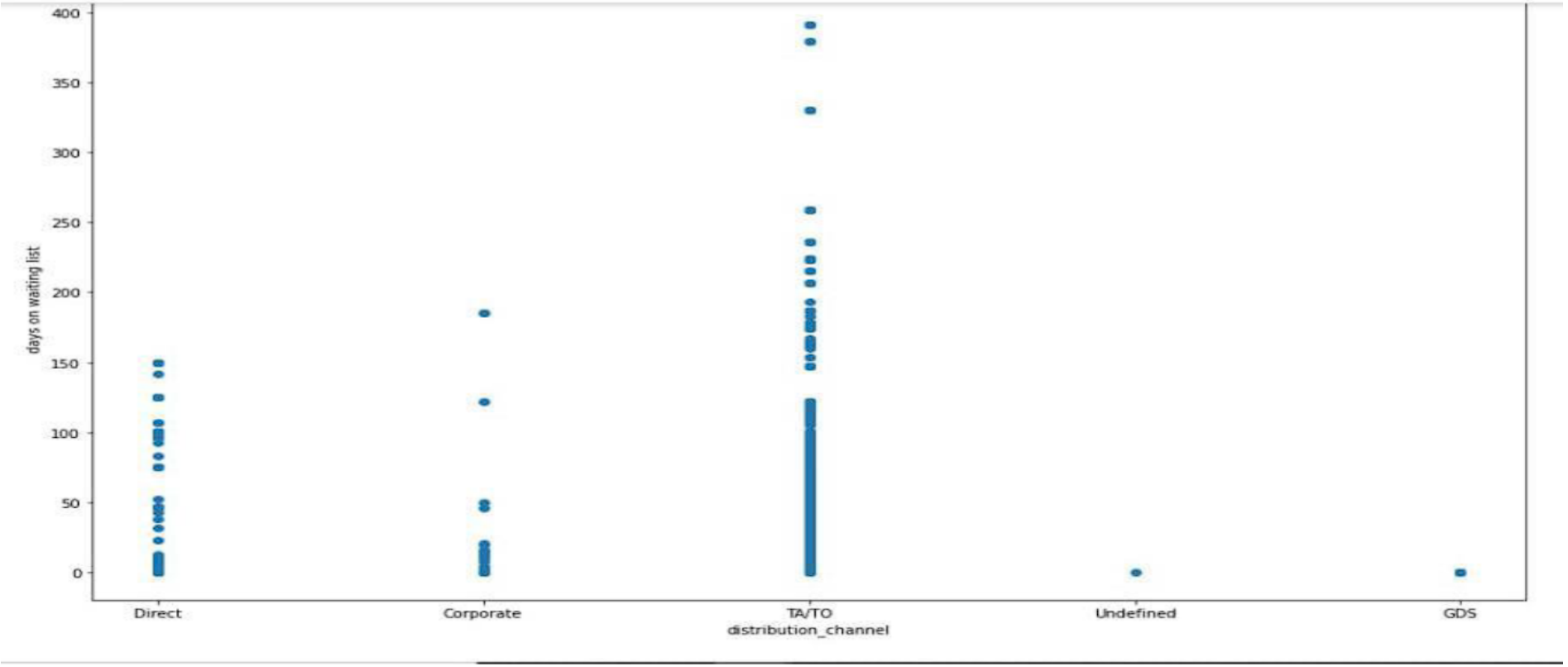
UNIVARIATE ANALYSIS

Number of Bookings for various types of hotels

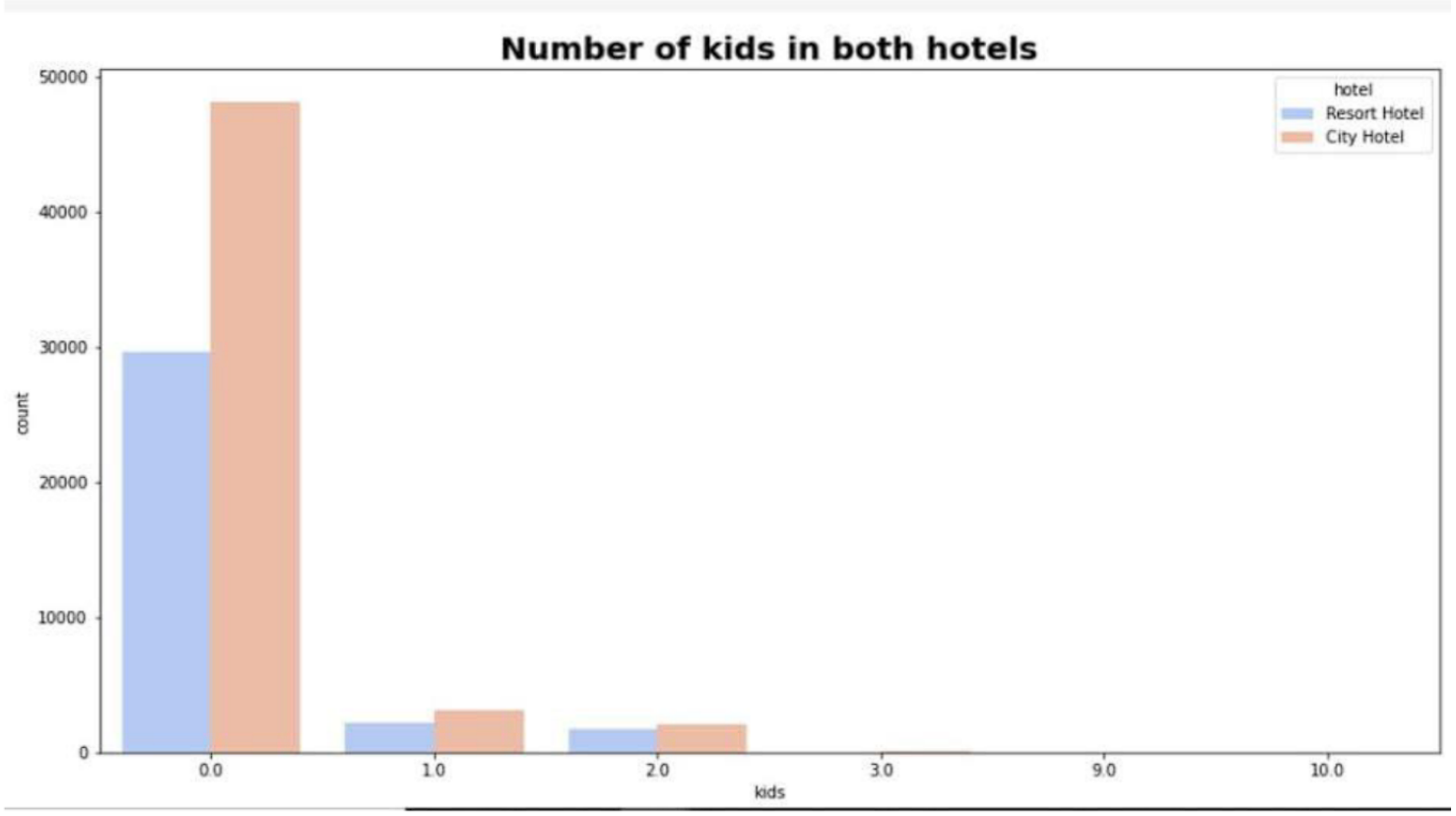
Plot between Type of market segment and Waiting list for **the** booking



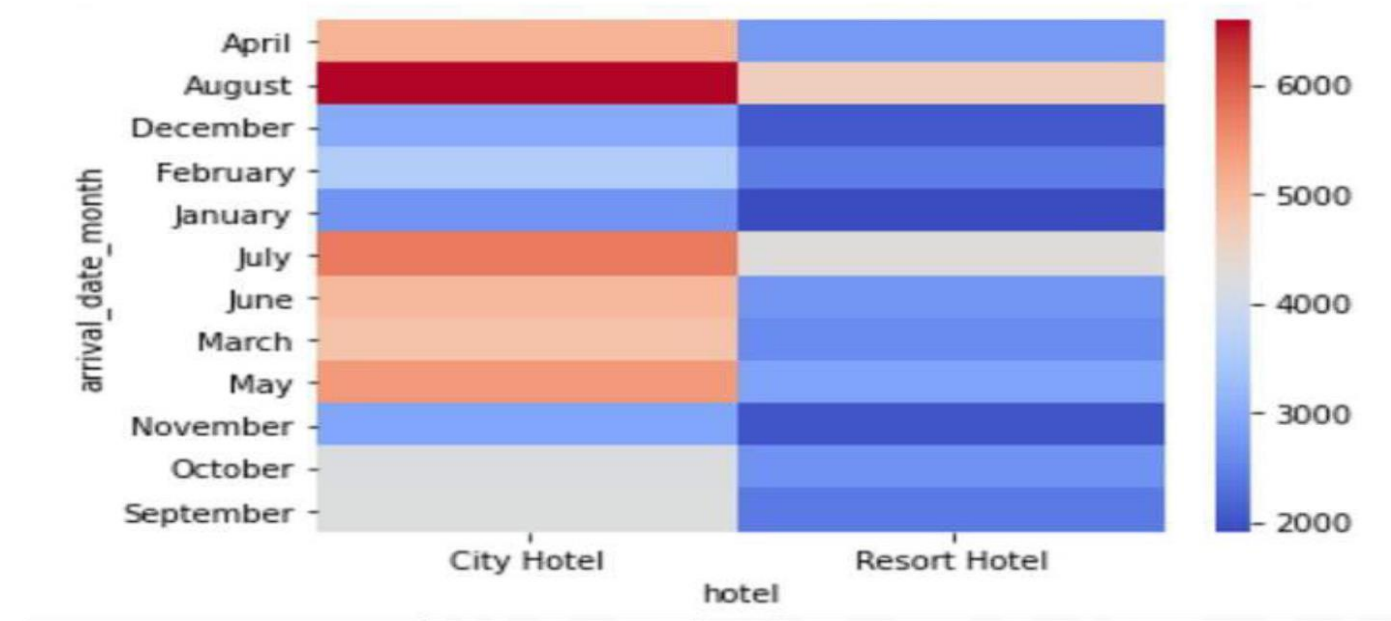
Plot between Distributing Channel and Days on the waiting



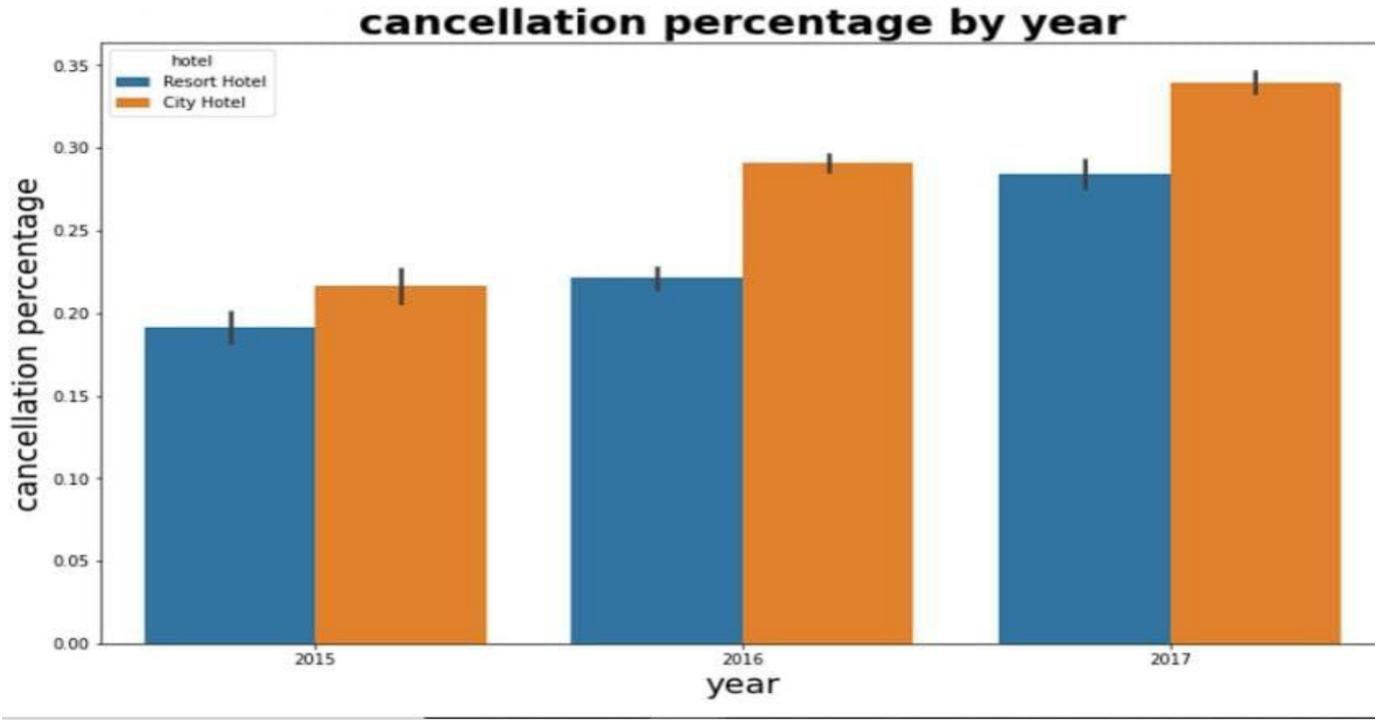
Number of kids in both hotels



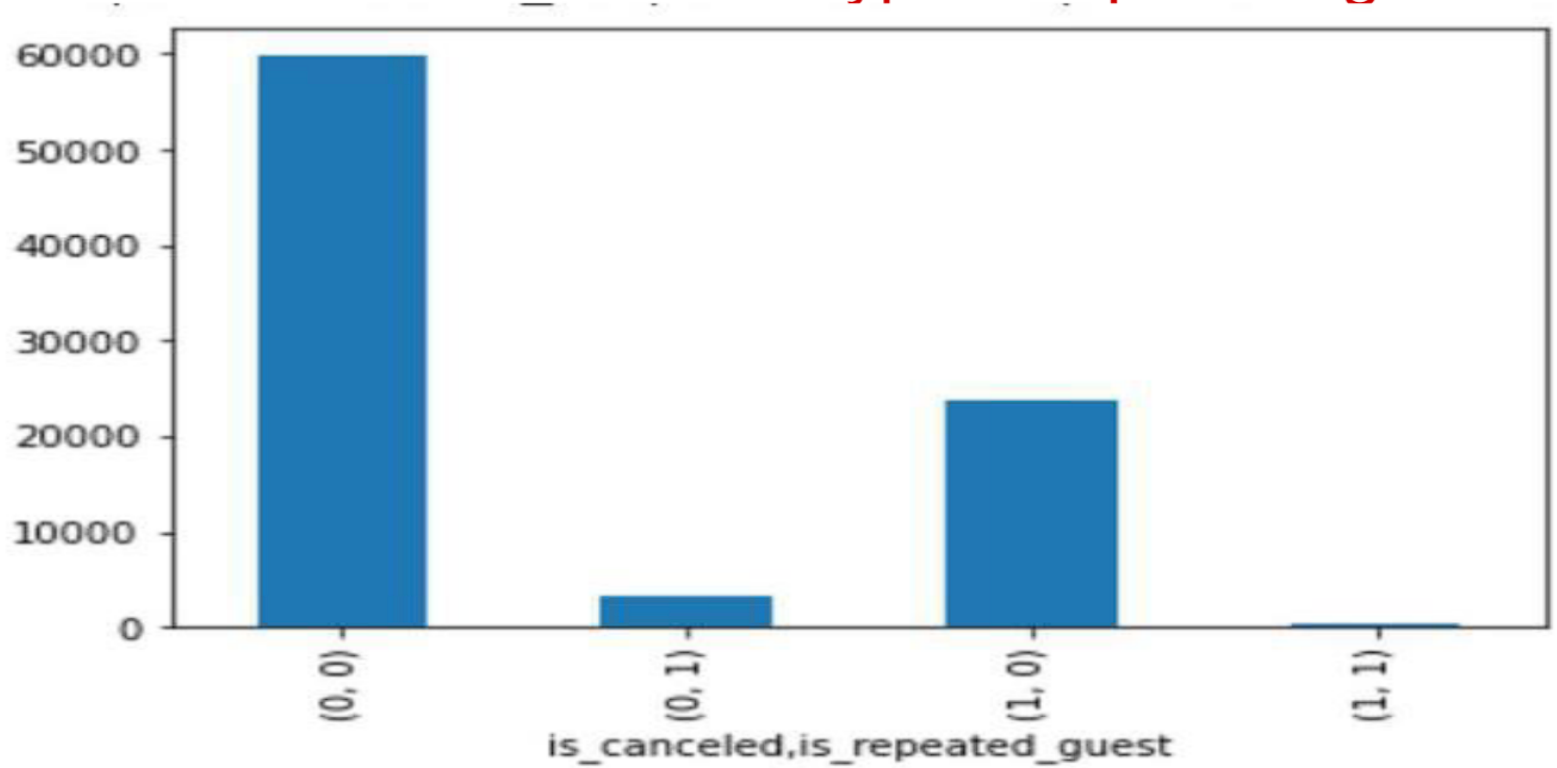
Heatmap between type of Hotel and arrival month



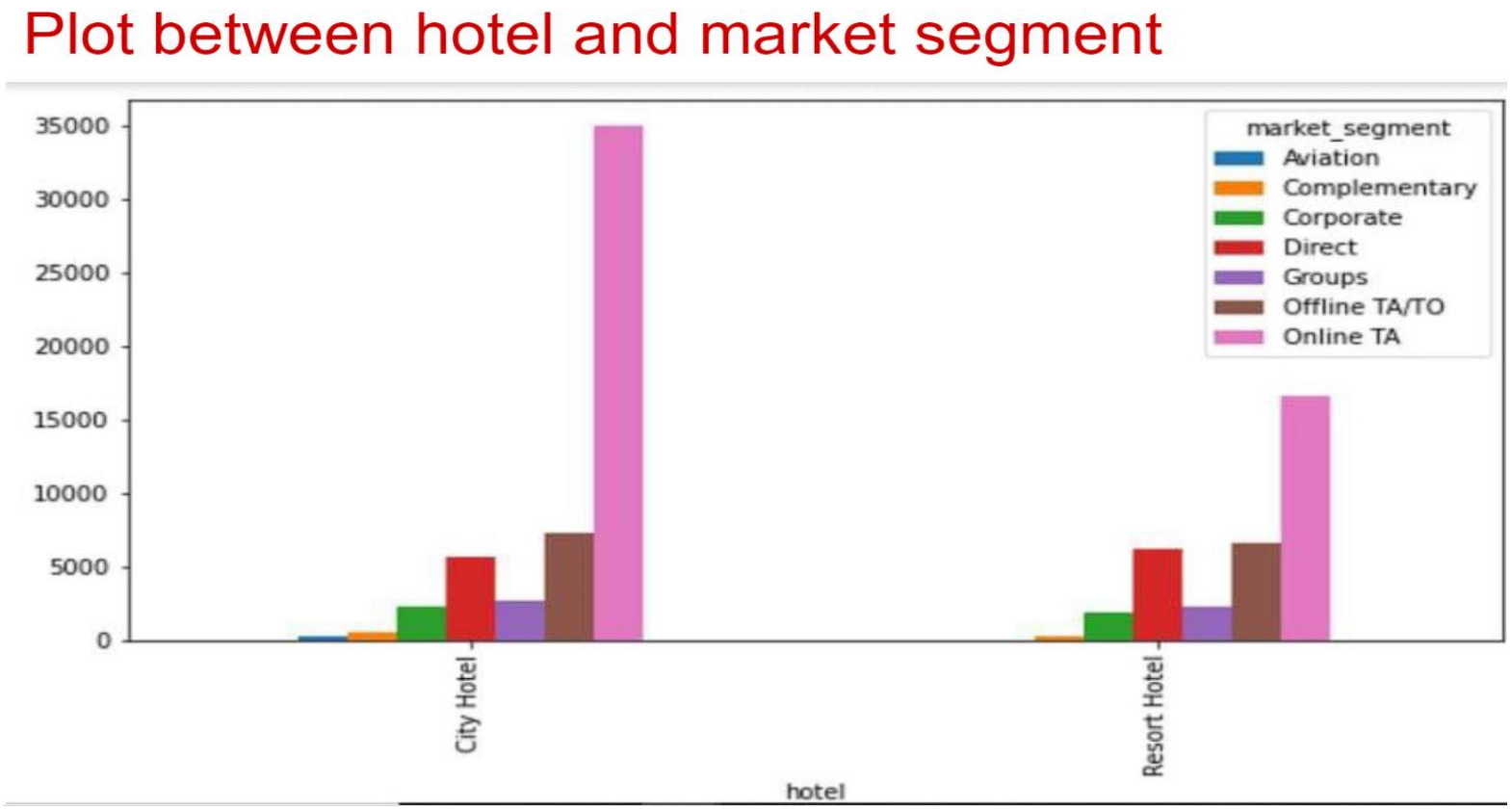
Cancellation percentage by year



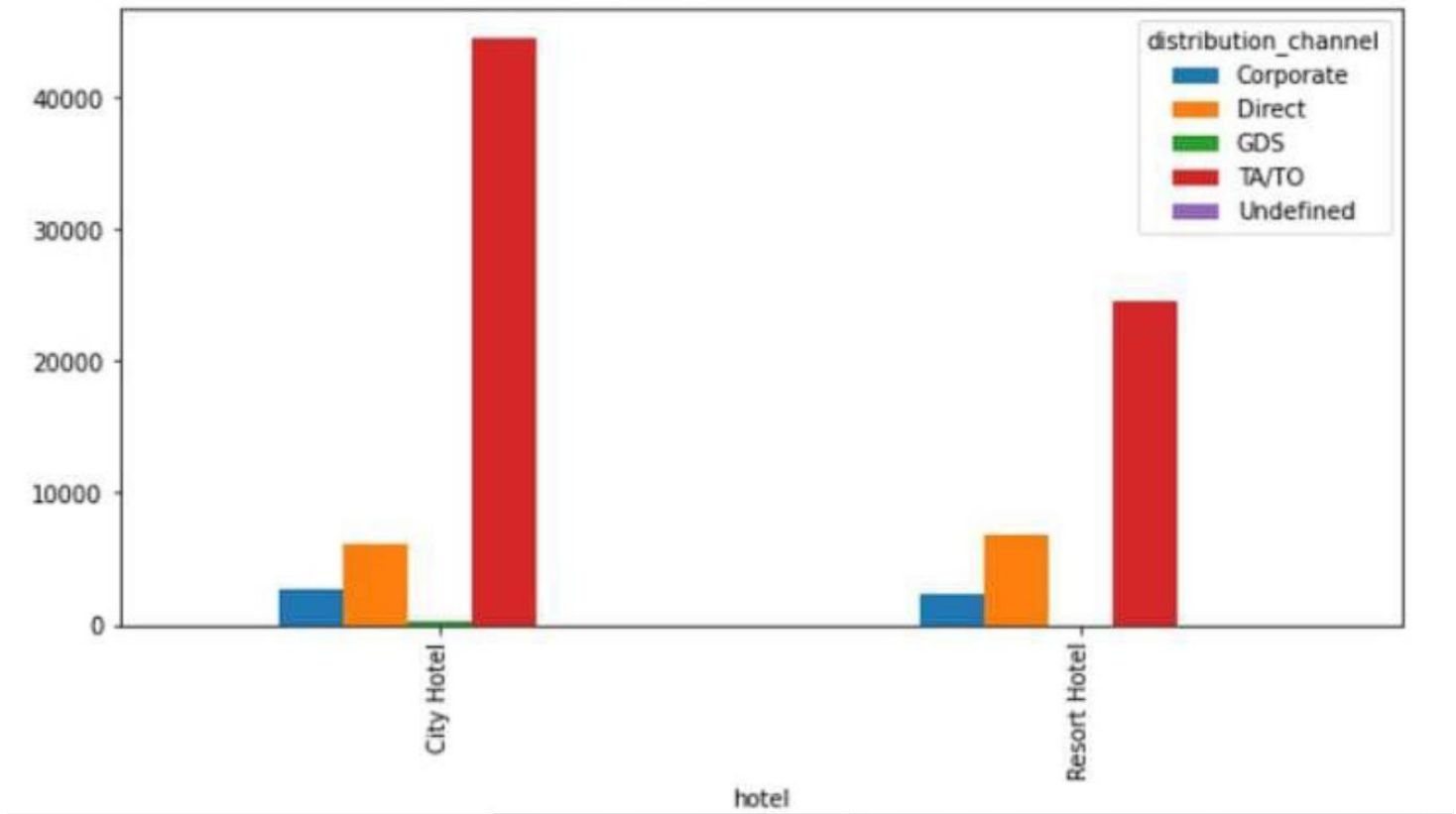
Plot between cancellation type & repeated guest



BIVARIATE AND MULTIVARIATE ANALYSIS

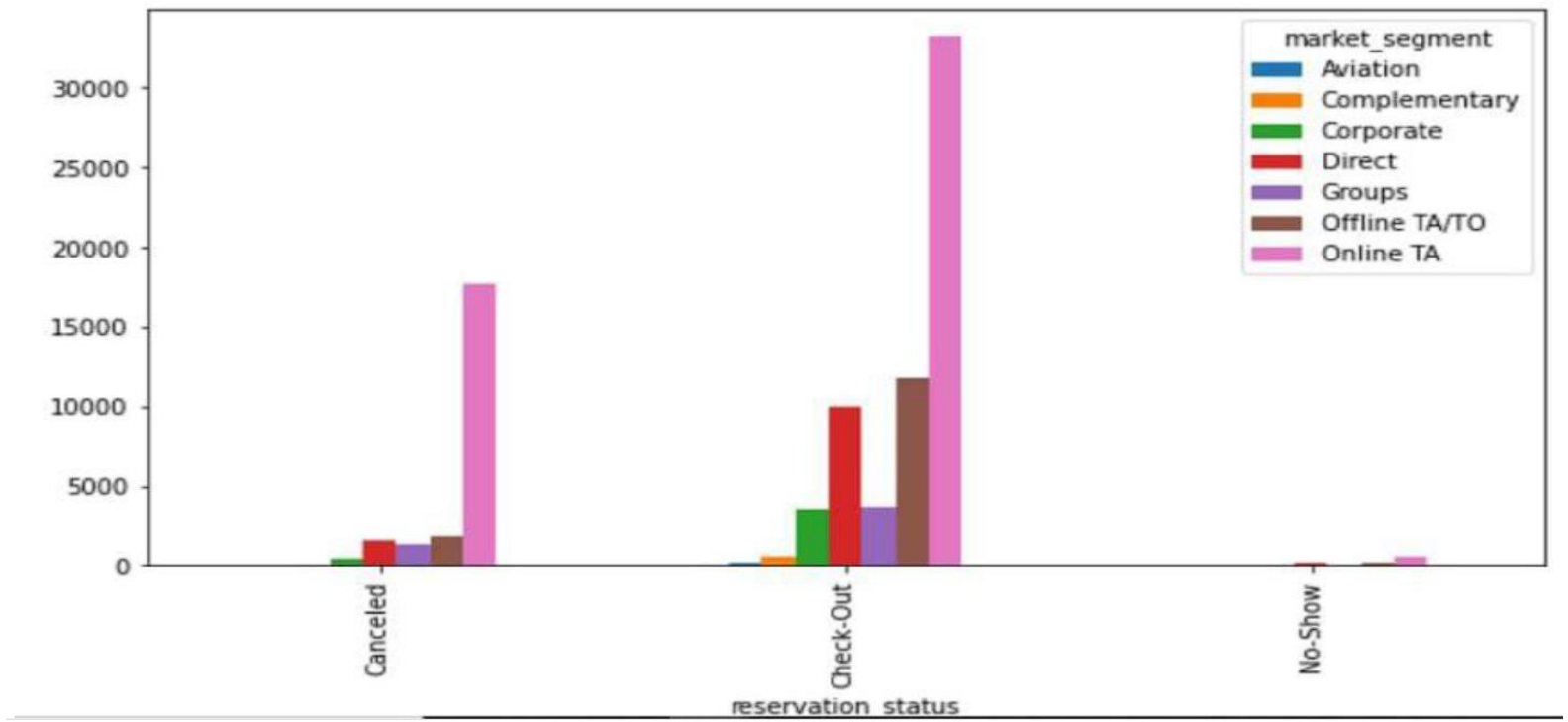


Plot between hotel and distribution channel

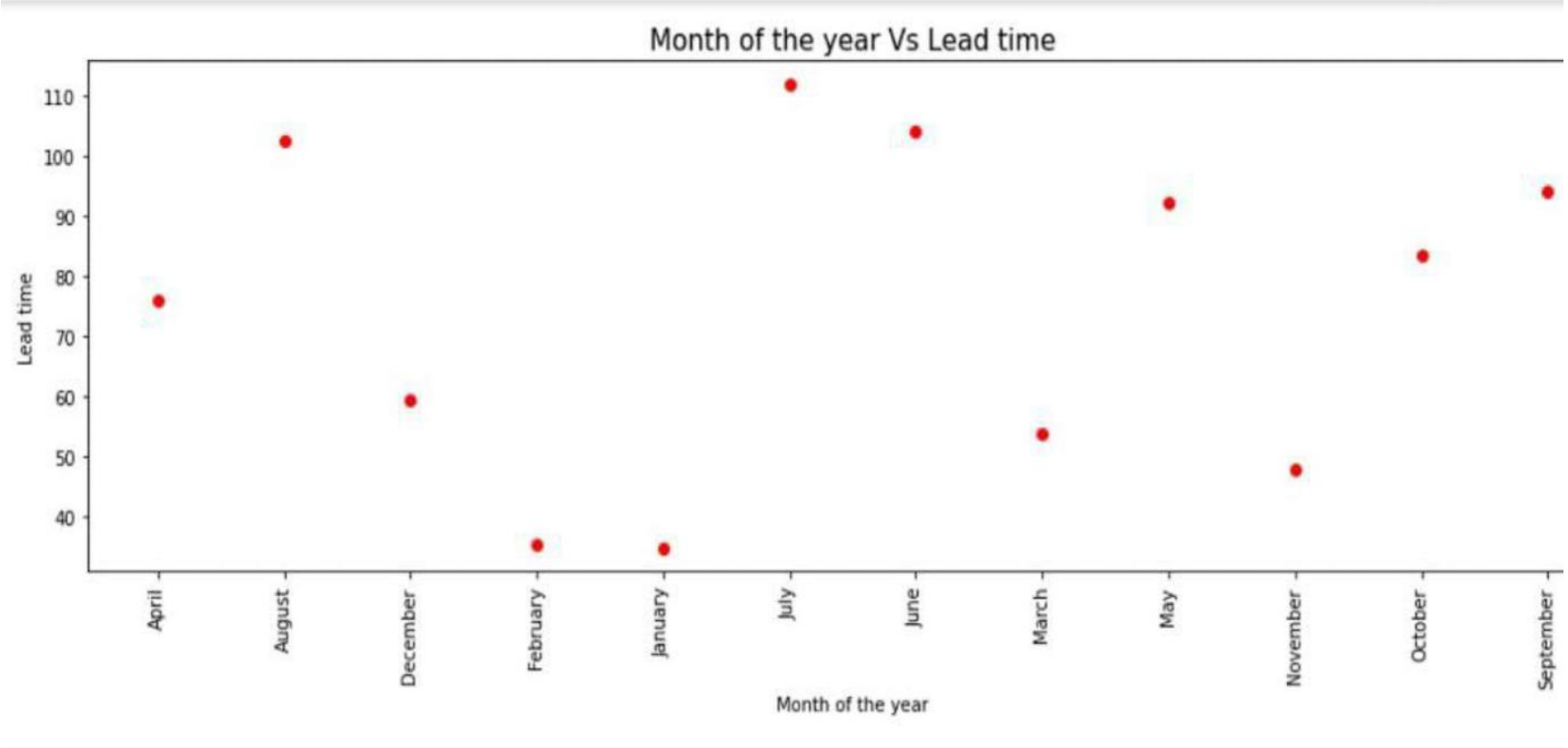


AlmaBetter

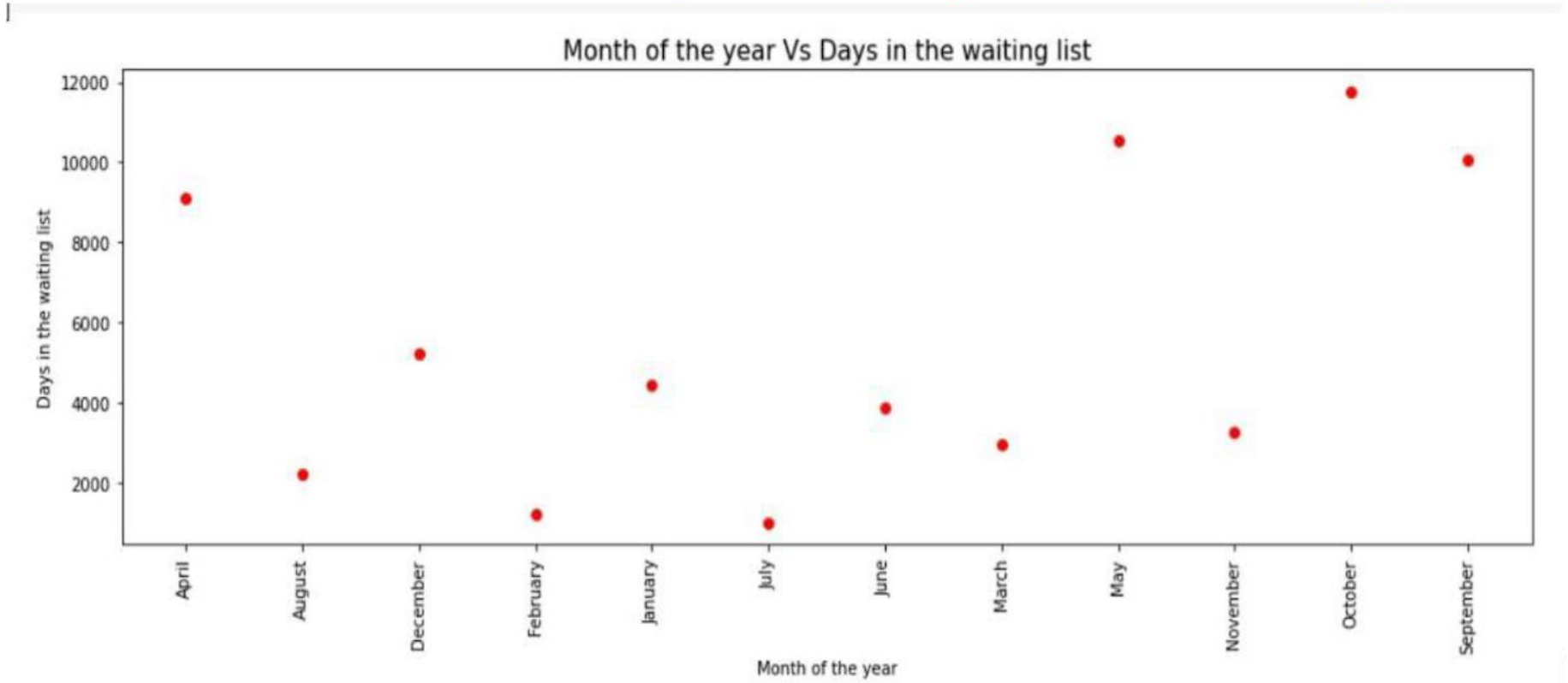
Plot between resevation status and market segment

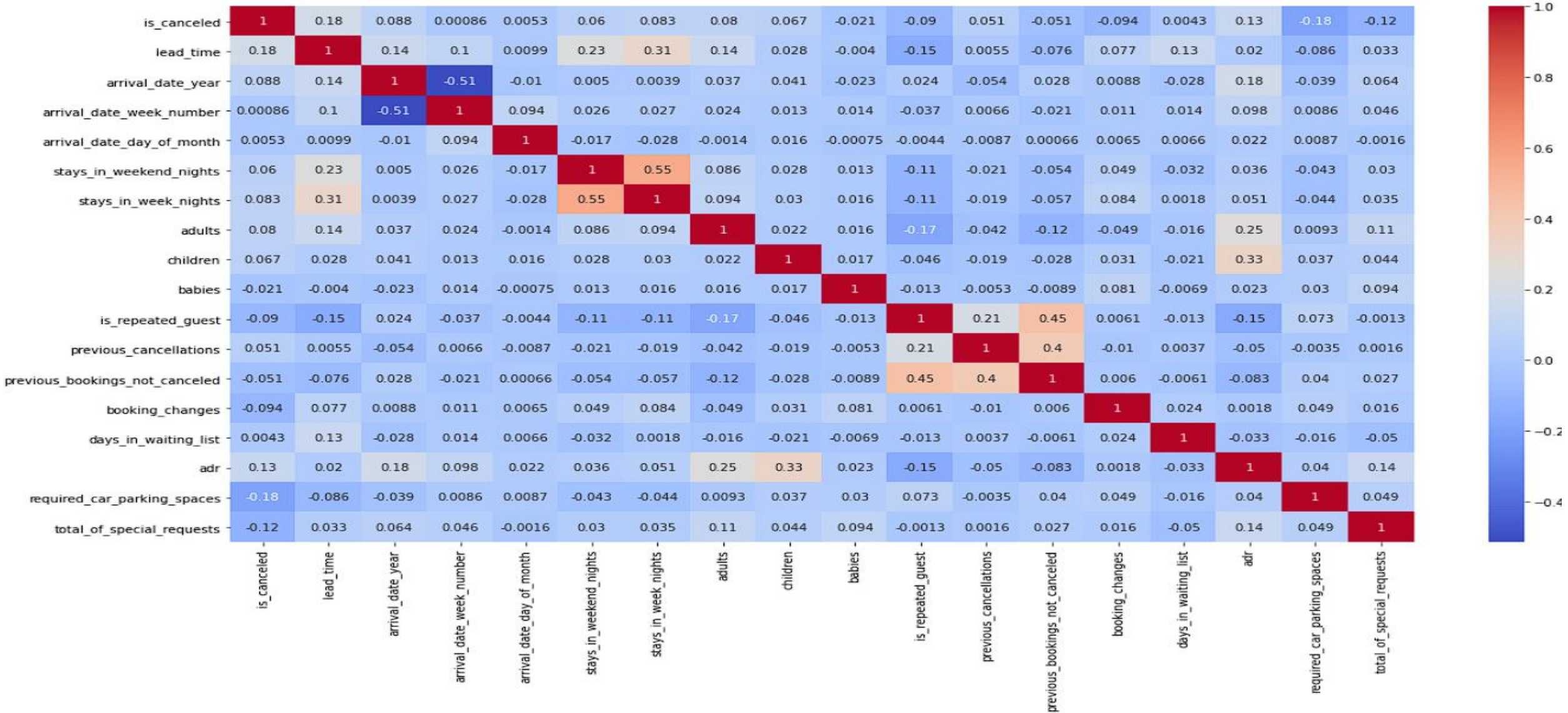


Plot between Month of year & Lead time



Plot between Month of year & Days in the waiting list




is_canceled 1 0.18 0.088 0.00086 0.0053 0.06 0.083 0.08 0.067 -0.021 -0.09 0.051 -0.051 -0.094 0.0043 0.13 -0.18 -0.12
lead_time 0.18 1 0.14 0.1 0.0099 0.23 0.31 0.14 0.028 -0.004 -0.15 0.0055 -0.076 0.077 0.13 0.02 -0.086 0.033
arrival_date_year 0.088 0.14 1 -0.51 -0.01 0.005 0.0039 0.037 0.041 -0.023 0.024 -0.054 0.028 0.0088 -0.028 0.18 -0.039 0.064
arrival_date_week_number 0.00086 0.1 -0.51 1 0.094 0.026 0.027 0.024 0.013 0.014 -0.037 0.0066 -0.021 0.011 0.014 0.098 0.0086 0.046
arrival_date_day_of_month 0.0053 0.0099 -0.01 0.094 1 -0.017 -0.028 -0.0014 0.016 -0.00075 -0.0044 -0.0087 0.00066 0.0065 0.0066 0.022 0.0087 -0.0016
stays_in_weekend_nights 0.06 0.23 0.005 0.026 -0.017 1 0.55 0.086 0.028 0.013 -0.11 -0.021 -0.054 0.049 -0.032 0.036 -0.043 0.03
stays_in_week_nights 0.083 0.31 0.0039 0.027 -0.028 0.55 1 0.094 0.03 0.016 -0.11 -0.019 -0.057 0.084 0.0018 0.051 -0.044 0.035
adults 0.08 0.14 0.037 0.024 -0.0014 0.086 0.094 1 0.022 0.016 -0.17 -0.042 -0.12 -0.049 -0.016 0.25 0.0093 0.11
children 0.067 0.028 0.041 0.013 0.016 0.028 0.03 0.022 1 0.017 -0.046 -0.019 -0.028 0.031 -0.021 0.33 0.037 0.044
babies -0.021 -0.004 -0.023 0.014 -0.00075 0.013 0.016 0.016 0.017 1 -0.013 -0.0053 -0.0089 0.081 -0.0069 0.023 0.03 0.094
is_repeated_guest -0.09 -0.15 0.024 -0.037 -0.0044 -0.11 -0.11 -0.17 -0.046 -0.013 1 0.21 0.45 0.0061 -0.013 -0.15 0.073 -0.0013
previous_cancellations 0.051 0.0055 -0.054 0.0066 -0.0087 -0.021 -0.019 -0.042 -0.019 -0.0053 0.21 1 0.4 -0.01 0.0037 -0.05 -0.0035 0.0016
previous_bookings_not_canceled -0.051 -0.076 0.028 -0.021 0.00066 -0.054 -0.057 -0.12 -0.028 -0.0089 0.45 0.4 1 0.006 -0.0061 -0.083 0.04 0.027
booking_changes -0.094 0.077 0.0088 0.011 0.0065 0.049 0.084 -0.049 0.031 0.081 0.0061 -0.01 0.006 1 0.024 0.0018 0.049 0.016
days_in_waiting_list 0.0043 0.13 -0.028 0.014 0.0066 -0.032 0.0018 -0.016 -0.021 -0.0069 -0.013 0.0037 -0.0061 0.024 1 -0.033 -0.016 -0.05
adr 0.13 0.02 0.18 0.098 0.022 0.036 0.051 0.25 0.33 0.023 -0.15 -0.05 -0.083 0.0018 -0.033 1 0.04 0.14
required_car_parking_spaces -0.18 -0.086 -0.039 0.0086 0.0087 -0.043 -0.044 0.0093 0.037 0.03 0.073 -0.0035 0.04 0.049 -0.016 0.04 1 0.049
total_of_special_requests -0.12 0.033 0.064 0.046 -0.0016 0.03 0.035 0.11 0.044 0.094 -0.0013 0.0016 0.027 0.016 -0.05 0.14 0.049 1
is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

CHALLENGES

- The name of the countries was not in the proper format,because of which we are not able to plot the geomap plot.

  Company and agent column has lots of duplicate value
- There were many rows with almost similar data
- Lots of null values in the dataset

CONCLUSION

- Month of August and july receives most no. of booking.
- Booking for city hotels is twice as for resort hotels.
- Repeated custumers cancel their hotel in very rare cases.
- Customers coming from aviation industry has very less time i.e. they book urgently
- People with no kid prefer to choose city hotel over resort hotel

Strategies to counter high cancellations at hotel

- Since we see, our repetitive costumers are most loyal costumers,to maintain them we can provide them with some bonus points,which can be redeem in the next booking
- Month of January and December receives less no. of booking,hotels can offer discounted packages for these months.
- Family with kids prefer resorts , we can provide with holiday family packages.
- Great no. of the bookings are coming from travel agents, so we can provide them some commission

THANK YOU